Statistical Applications in Genetics and Molecular Biology

Volume 3, *Issue* 1 2004 *Article* 22

A Mixed Model Approach to Identify Yeast Transcriptional Regulatory Motifs via Microarray Experiments

Xiang Yu, *Bioinformatics Research Center, North Carolina State University; Merck & Co., Inc., RY80M-180C, P.O. Box 2000, Rahway, NJ 07065-0900*
Tzu-Ming Chu, *Department of Genomics, SAS Institute Inc*
Greg Gibson, *Bioinformatics Research Center, North Carolina State University; Department of Genetics, North Carolina State University*
Russell D. Wolfinger, *Department of Genomics, SAS Institute Inc*

[CORRECTED VERSION]

Recommended Citation:
Yu, Xiang; Chu, Tzu-Ming; Gibson, Greg; and Wolfinger, Russell D. (2004) "A Mixed Model Approach to Identify Yeast Transcriptional Regulatory Motifs via Microarray Experiments," *Statistical Applications in Genetics and Molecular Biology*: Vol. 3: Iss. 1, Article 22.
DOI: 10.2202/1544-6115.1045
Available at: http://www.bepress.com/sagmb/vol3/iss1/art22

A Mixed Model Approach to Identify Yeast Transcriptional Regulatory Motifs via Microarray Experiments

Xiang Yu, Tzu-Ming Chu, Greg Gibson, and Russell D. Wolfinger

[Corrected Version]

Abstract

A genome-wide location analysis method has been introduced as a means to simultaneously study protein-DNA binding interactions for a large number of genes on a microarray platform. Identification of interactions between transcription factors (TF) and genes provide insight into the mechanisms that regulate a variety of cellular responses. Drawing proper inferences from the experimental data is key to finding statistically significant TF-gene binding interactions. We describe how the analysis and interpretation of genome-wide location data can be fit into a traditional statistical modeling framework that considers the data across all arrays and formulizes appropriate hypothesis tests. The approach is illustrated with data from a yeast transcription factor binding experiment that illustrates how identified TF-gene interactions can enhance initial exploration of transcriptional regulatory networks. Examples of five kinds of transcriptional regulatory structure are also demonstrated. Some stark differences with previously published results are explored.

Author Notes: We thank R.A. Young's lab for publishing their experimental data online, N.J. Rinaldi and B.S. Weir for discussions and comments on the manuscript. A previously published version of this paper had an error in the normalization model. This version has the corrected results; the previous version is stored for archival purposes and is available at the same web address.

Introduction

The functionality of a living cell is realized through the concerted activity of a set of genes and gene products. Genome-wide transcriptional analysis, for example, with microarray techniques, provides important information about the expression profiles of clusters of genes that are involved in a variety of cellular processes (DeRisi *et al*., 1997; Cho *et al*., 1998; Gasch *et al*., 2000), yet the how these genes are coordinately regulated remains a topic of intense investigation. Recently, probabilistic models were introduced to build regulation modules using gene expression data from a collection of yeast microarray experiments that measure transcriptional responses to a variety of stress conditions (Segal *et al*., 2001; Pe'er *et al*., 2001; Segal *et al*., 2003). The algorithm developed by Segal *et al*. (2003) takes gene expression data and a large precompiled set of candidate regulatory genes including transcription factors and intermediate signal transduction molecules. Genes are clustered by expression profile based on a regression tree in which each internal node specifies a regulatory input and each leaf node contains a set of genes that respond to a series of regulatory contexts defined by the path from the root node to that leaf node.

With their ability to selectively recognize and bind the promoter of specific genes, transcription factors play a key role in controlling gene expression. Using gene expression profiles, however, whether a transcription factor specifically activates its target genes is usually inferred indirectly from the transcription level of that transcription factors *per se,* as well as the expression patterns of the target genes. Detecting the binding interactions between transcription factors and genes provides extra information that will help to reveal functional features of the genome, and may shed light on the regulatory mechanisms behind complex cellular process that are coordinated by programmed gene expression involving multiple genes and regulators.

A conventional method to detect protein-DNA interaction *in vivo* is chromatin immunoprecipitation (IP), in which DNA fragments cross-linked to a protein of interest are enriched with a specific antibody to that protein (Orlando, 2000). The introduction of microarray technology has enabled investigators to study simultaneous changes in expression across a large number of genes. Although they have been widely used in assessing differentially expressed genes under different conditions, microarray experiments are not limited to expression profiling. By integrating a modified chromatin immunoprecipitation procedure into DNA microarray analysis, Ren *et al*. (2000) developed a genome-wide location analysis method, and demonstrated its capacity to monitor protein-DNA interaction at a genome-wide scale. In their experiments, chromatin immunoprecipitation was used to extract DNA bound by a transcription factor, and IP-enriched DNA fragments (representing promoter sequences bound by that

transcription factor) and un-enriched ones were labeled with different fluorophores and hybridized to a microarray containing all known yeast gene promoter sequences. Different fluorescence intensities in the two dyes revealed the presence or absence of binding of that transcription factor to the gene promoter. A recent study of transcriptional regulation in *Sacchromyces cerevisiae* (Lee *et al*., 2002) adopted this genome-wide location analysis method to systematically examine the interactions between 106 known transcription factors and over 6,000 genes. They identified about 4,000 significant binding interactions, which not only helped them to define which transcription factors regulate which genes, but also implied possible structures for pieces of the transcriptional regulatory network.

With a collection of tens of thousands of observations from these microarray experiments, statistical interpretation and inferences from the data becomes the next challenge. Various methods have been developed to infer differentially expressed genes from expression microarray data, and these can also be applied to genome-wide location analysis data. Of these methods, ANOVA-type approaches (Kerr *et al*., 2000; Wolfinger *et al*. 2001; Chu *et al*., 2002) have been advocated for their superiority in partitioning sources of variation, providing reliable estimates of error variance, and for their flexibility in accommodating various experimental designs. Here we describe how an advanced ANOVA model known as a mixed model can be appropriately set up to analyze the genome-wide location data from Lee *et al*. (2002). For each gene, a mixed model is fitted that uses all observations across all 106 transcription factors for that gene, and a testable hypothesis concerning whether a transcription factor binds to the promoter of that gene is constructed. Significant TF-gene binding is then used to build up transcriptional regulatory motifs that may provide insights as to how transcription factors coordinate in regulation of a set of genes in response to specific cellular processes.

Experimental design and data

In their study, Lee *et al*. (2002) constructed, for each known yeast transcription factor (TF), a strain in which that transcription factor is tagged with a *c-myc* epitope, resulting in 106 tagged strains. Immunoprecipitation was then used to separate DNA fragments representing promoter regions cross-linked to the transcription factor using an anti-Myc antibody. After separation, two pools of DNA, one bound by the transcription factor and the other not, were labeled with different dyes (Cy3 or Cy5) and hybridized to a DNA microarray on which promoter regions of over 6,000 yeast genes were spotted. TF-promoter interactions were then identified by increased fluorescence intensity in the dye coupled to the IP-enriched DNA pool. Three replicates were done for each

DOI: 10.2202/1544-6115.1045

transcription factor and final data were obtained from 300 arrays. The images of the arrays were processed by either ScanAlyze® or ArrayVision® and the raw data files can be downloaded from the website of Young's lab.

Data quality inspection

Prior to any analysis, it is important to check the data quality and filter aberrant

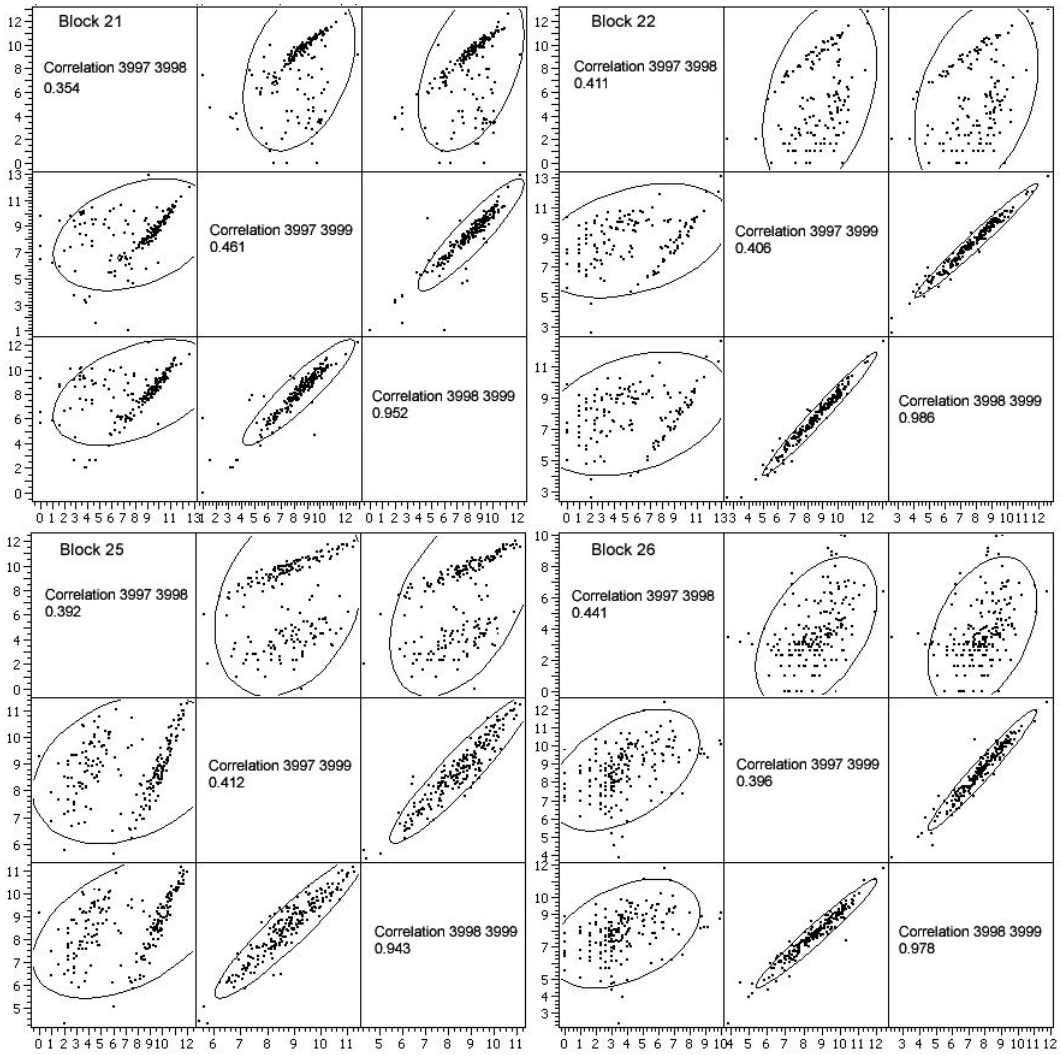

Figure 1. Scatter plot of block 21, 22, 25 and 26 of array 3997 for transcription factor Zap1 with its two replicates (array 3998 and array 3999). The four panels represent block 21 (top left), 22 (top right), 25 (bottom left), 26 (bottom right), respectively and the eclipse in each cell represents 90% density curve.

observations. A simple yet effective way is to check the correlations between replicates. We created scatter plots between replicates for each of the 106 transcription factors and found 7 out of 300 arrays that show low correlations ($\leq$ 0.6) with their corresponding replicates. The seven arrays are one of the replicates from seven different experiments with transcription factors Fkh2, Mcm1, Mot3,

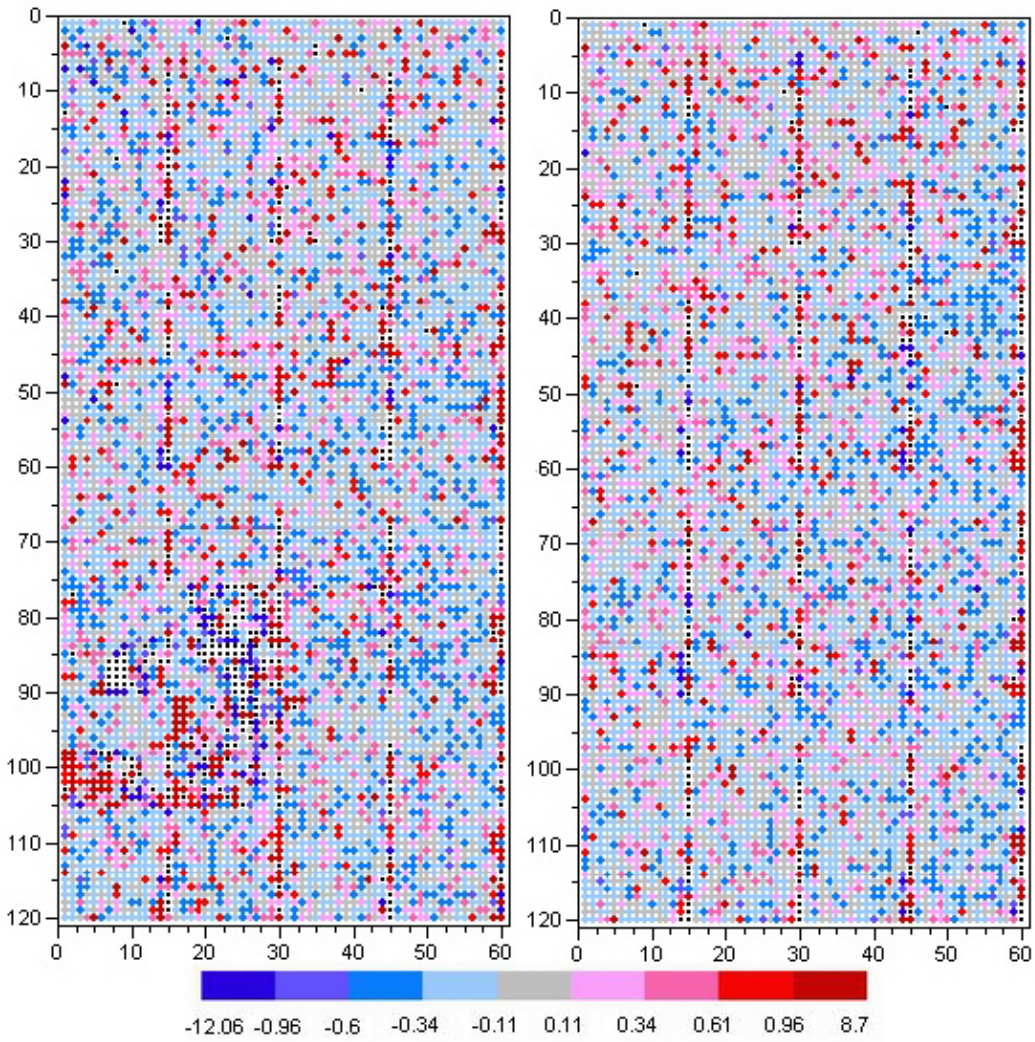

Figure 2. Pseudo image of array 3997 and its replicate 3998 from transcription factor Zap1. In a pseudo image, the raw intensity ratios are converted to RGB values from 0 to 255 and the positions of each spot are used as coordinates to plot that spot along with its RGB value. The image is composed of 8×4 = 32 blocks each with 15×15 = 225 spots. The *X* axis ranges from 0 to 60 and the *Y* axis ranges from 0 to 120 and for each spot (x, y), its block identifier is determined by $2^{ceil(y/15)}+ceil(x/15)$. It can be seen from the pseudo image that the area containing spots with coordinates $0 \leq x \leq 30$ and $75 \leq y \leq 105$ shows apparent difference compared to its replicate, which corresponds to the four blocks 21, 22, 25 and 26.

DOI: 10.2202/1544-6115.1045

Rph1, Swi5, Swi6 and Zap1. A closer look at the dubious array for transcription factor Rph1 showed that over 75% of intensity values are below 11, a number that is normally considered to correspond to background. As a comparison, the 25% percentiles of its two replicates are 54 and 60, and the corresponding 75% percentiles are 281 and 366, respectively. We decided to discard this array since its quality was too low, possibly due to failure of the hybridization. An important characteristic of a typical cDNA microarray is that the array is prepared in blocks, because the DNA is typically spotted on the slide sequentially with a 4-pin or 16-pin printer. The recorded block identifier makes it possible to track the physical location of each spot and detect local contamination in a small region. The latter can be done using a scatter plot for each block between replicates. We show in Figure 1 that, in the array from the problematic transcription factor Zap1, 4 blocks out of 32 exhibit almost no correlation with their replicates while other blocks correlate well. A pseudo image (Figure 2) was created that maps the RGB-converted intensities to the physical locations of each spot for this array as well as one of its replicates. The image indicated likely local contamination in a region containing mainly these four blocks. After removal of the observations in these four blocks, the correlation between this array and its two replicates was significantly improved (Figure 3). For the other five arrays, we did not detect significant block effects and concluded that the low correlations come from experiment-wise variations and kept all those observations in our analysis that follows. The cleaned data set contains 299 arrays with promoters of 6,279 genes each, except for one array from Zap1 in which 900 observations in the four questionable blocks were eliminated.

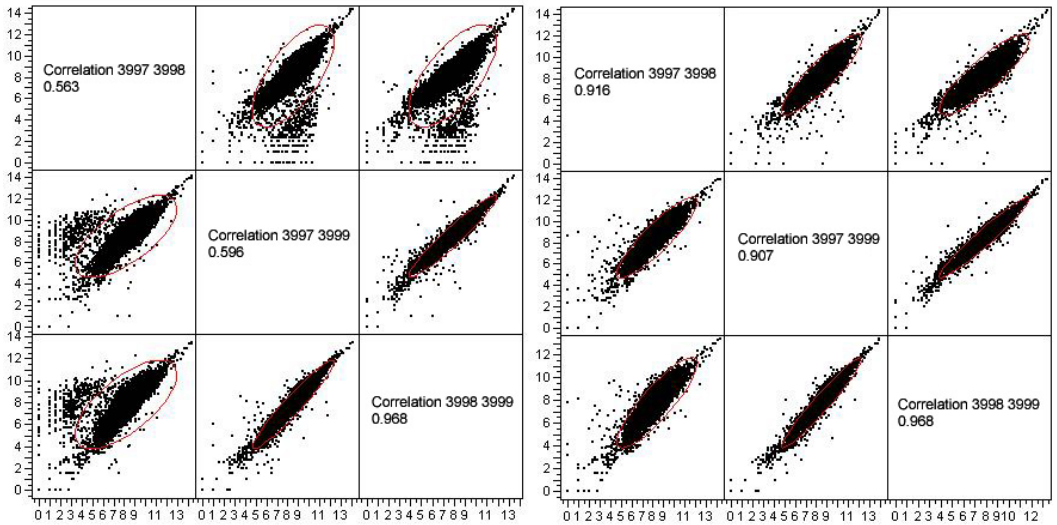

Figure 3. Scatter plot of array 3997 with its two replicates (array 3998 and array 3999) before (left panel) and after (right panel) removal of dubious blocks 21, 22, 25 and 26.

The error model

In their analysis, Lee e*t al*. (2002) used an error model following from Hughes *et al*. (2001). The error model assumes that the error term is a combination of not only an additive component that is estimated from the individual observations for each particular gene, but also a fractional multiplicative component that is derived empirically and tracks the increased variation towards low intensities. Combining these two error components leads to a conservative estimate of signal variability and as the number of replicates increases, the modeled, multiplicative error decreases. This combined error is used with the log intensity ratio from the two channels to determine whether the promoter of a gene is significantly bound by a transcription factor.

Mixed model normalization and analysis

As an alternative analysis approach, we employ a two-stage linear mixed model to log transformed data, as described in Wolfinger *et al*. (2001). Let y_{gtda} be the base-2 logarithm of the background-corrected intensity from gene g (g = 1, …, 6279), transcription factor t (t = 1, …, 106), labeled with dye d (d = 1 for Cy3 and d = 2 for Cy5) in array a (a = 1, …, 299). The normalization model is:

$$y_{gtda} = \mu + T_t + D_d + A_a + (TD)_{td} + (DA)_{da} + \xi_{gtda},$$

where μ represents the global mean value, T is the main effect of transcription factor, D is the main effect of dye, A is the random effect for arrays, and TD and DA are the factor-by-dye and dye-by-array interactions, respectively. T, D and TD are assumed to be fixed, A, DA and ξ are assumed to be normally distributed random variables with mean 0 and variance σ_A^2, σ_{DA}^2 and σ_e^2, respectively. This normalization corrects for effects across the entire arrays, that is, are not gene specific. Residuals from the normalization model are then used as the input to our second mixed model, in which we fit each gene individually:

$$r_{tpa} = \mu + T_t + P_p + (TP)_{tp} + A_a + \varepsilon_{tpa}.$$

Here, r_{tpa} is the residual measurement for transcription factor t (t = 1, …, 106), probe p (p = 1 for IP-enriched DNA pool and p = 2 for genomic control DNA) and array a (a = 1, …, 299). μ is the global mean, T is the main effect of transcription factor representing the overall differences in intensity for each yeast strain in which a different TF is tagged, P is the probe main effect representing the overall differences in intensity for two DNA pools hybridized to the array, and TP is the effect of transcription factor by probe interaction representing the intensity differences between the two probes within the same transcription factor. In these gene models, this interaction term $(TP)_{tp}$ accounts for the gene specific

DOI: 10.2202/1544-6115.1045

differences of binding intensities between transcription factor *t* and the genomic control. *A* is the array random effect representing overall variation in fluorescence

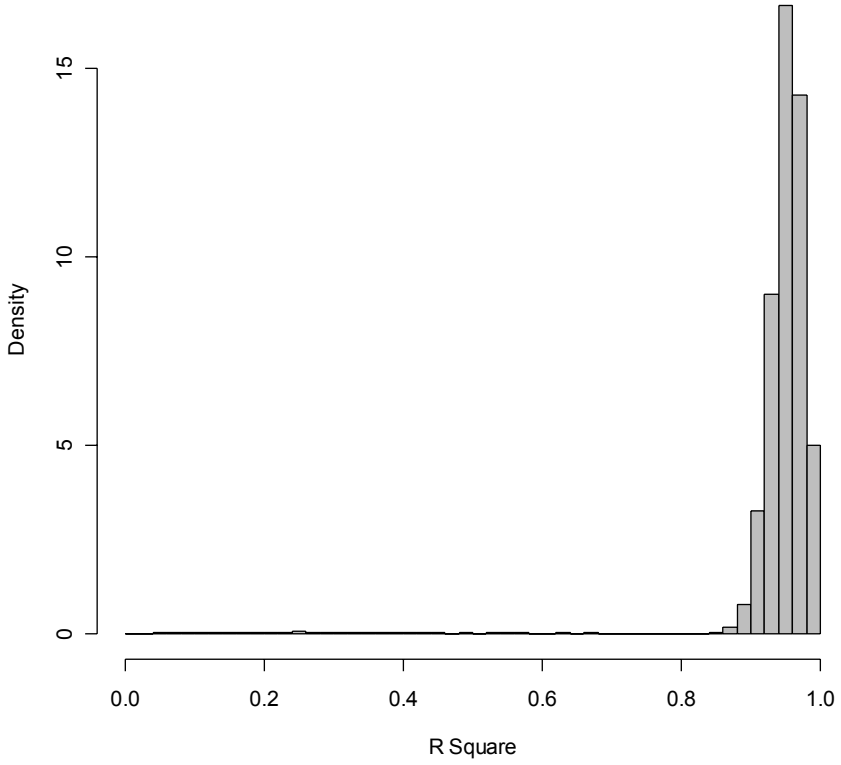

Figure 4. Histogram of R^2 values for the 6,279 genes from the mixed model fitting.

measurement across arrays and is assumed to be normally distributed with mean 0 and variance σ_A^2. ε is the error term and is assumed to be normally distributed with mean 0 and variance σ^2 and is independent of *A*. Typically, a dye effect term that corrects for gene specific intensity differences between the two different dyes needs to be added to the gene model. In this experiment, all IP probes were labeled with one dye (Cy5) and genomic control probes labeled with another (Cy3), except for 35 out of 300 arrays where the dyes are reversed, that is, IP probe was labeled with Cy3 and control probe was labeled with Cy5. However, within each transcription factor, all three replicates were labeled in one direction without dye-flip. Therefore, the dye effect is completely confounded with some combinations of probe and factor effects so it is not estimable.

We also perform a simple iterative outlier filtration method based on mixed model fitting. For each iteration, we first fit the mixed model and calculated the standardized residual for each observation, then eliminate observations having a standardized residual greater than three (Chu *et al.*, 2002). A more sophisticated method is to perform a statistical test for each potential outlier and filter outliers one-by-one (Chu and Wolfinger, 2003) but generally an observation with a standardized residual larger than three can be regarded as an outlier based on the assumption of normally distributed errors. We carry out three iterations and filtered ~0.30% observations as outliers. Data after outlier screening are then fit by our mixed model to obtain the final results. The median R^2 values of the 6,279 gene models, as shown in Figure 4, was about 0.95, indicating that the mixed model offers a good fit to the data and explains most of the observed variation.

As described, our goal is to find, for each gene, if a transcription factor *t* that significantly binds the promoter of the gene, as indicated by increased fluorescence intensity in IP-enriched probe. Therefore, to test the hypothesis whether a transcription factor *t* interacts with the promoter of a gene, we need to contrast the transcription factor by probe interaction term between the IP-enriched probe and the genomic control probe. Our hypothesis test is then formulized as:

$$H_0 : \ TP_{t1} - TP_{t2} <= 0,$$

$$H_1 : \ TP_{t1} - TP_{t2} > 0,$$

and a one-sided *t*-test is used to assess the significance for each of the 6,279 genes and 106 transcription factors.

Using a Bonferroni correction for these 6,279×106 (665,574) tests, we find 12,147 significant TF-gene interactions at the 0.05 level, which corresponds to roughly 8 false positives out of 100 million single tests. Of the 6,279 genes, 5,183 have been fully annotated and named with known ORFs from the *Sacchromyces cerevisiae* genome database (SGD, 2003). The 12,147 TF-gene interactions we found involve 104 transcription factors and 3,608 genes. Since the Bonferroni adjustment is usually thought to be too conservative, we also performed a false discovery rate (FDR) adjustment (Benjamini and Hochberg, 1995) for the *p*-values obtained from the over 600K tests. For example, the FDR adjusted *p*-value at 0.01 is equivalent to a *p*-value of 6.3×10^{-4} in the original test, which results in 40,386 significant TF-gene interactions. As a comparison, the error model (Stoughton and Dai, 2002) identified 3,985 significant interactions using a single test *p*-value threshold of 0.001. A *p*-value cutoff of 0.05 from the error model would have identified 35,365 significant interactions. Using the same Bonferroni *p*-value cutoff of 7.5×10^{-8}, however, only 427 significant interactions would be retained from the error model. Results from different *p*-value cutoffs for multiple-testing adjustment are summarized in Table 1. The correlation of the *p*-value from the mixed model and that from the error model is 0.45 and the *p*-value rank correlation of the two models is also 0.45, indicating a reasonable consistency of

DOI: 10.2202/1544-6115.1045

the two analyses. To check whether the data filtering process prior to the analysis has a major impact on the results obtained, we repeated the same analysis using the original, un-cleaned data and the corresponding *p*-value and rank correlations are still 0.45 with slight changes at the third decimal. The number of significant bindings identified using the un-cleaned data is 12,074. This indicates a slight improvement of the analysis by filtering bad spots in advance, but the majority of

Table 1. Comparison of multiple-testing adjustment for the mixed model and the error model. The number of significances using the nominal ($\alpha = 0.05$), false discovery rate (FDR) and Bonferroni thresholds are recorded for the two models. The total number of tests is 665,574.

Method	Cutoff	Single test *p*-value	Number of Significances	
			Mixed Model	Error Model
Nominal	.05	.05	107,924	35,365
FDR	.01	6.3×10^{-4}	40,386	3,190
FDR	.001	4.2×10^{-5}	26,506	1,380
FDR	.0001	2.9×10^{-6}	18,678	821
Bonferroni	.05	7.5×10^{-8}	12,147	427

the results do not change because the proportion of outliers is relatively small to the large volume of the data.

As a comparison of the results from the mixed model and the error model, the negative base-10 log *p*-value of the 6,279×106 (665,574) tests from the two models using the un-cleaned data are plotted against each other in Figure 5. The horizontal dashed line indicates the Bonferroni cutoff (*p*-value = 7.5×10^{-8}) used in the mixed model and the vertical dashed line indicates the cutoff (*p*-value = 0.001) Lee *et al.* (2002) used from their error model. The upper right rectangular section contains significant test results found by both models whereas the upper left rectangular section represents significances identified by the mixed model but not by the error model and the lower right section indicates the opposite. The counts in the above three sections are 2,114, 10,033, and 1,871, respectively.

To further investigate the differences between these two models, we checked the mean difference of the log base-2 intensity between the IP probe and the control probe (log-ratio of the two channels) for all the positives identified by either of the two models. The results are shown in Figure 6. We can see that significant bindings found by both models are concentrated at high intensities with strong signals in the IP channel (log-ratio > 1 or 2-fold change at the raw intensity). The error model identified interactions with weaker signals that may be false negatives from the mixed model. These "false negatives" could be reduced with variance shrinkage methods that combine information across genes to obtain variance estimates (Lönnstedt and Speed, 2002; Feng *et al*., 2004; Cui *et al*., 2005). However, it is surprising that most of the "false negatives" from the error model

that are claimed positive by the mixed model show apparent two or more fold change (log-ratio > 1) at the raw intensity scale, which are highly likely to be true positives. Since there are only 3 replicates for each transcription factor in this experiment, the intensity-dependent, multiplicative error estimate from the error model may dominate the combined error, especially for low intensity spots, which may lead to inflated variances and thus reduce the sensitivity.

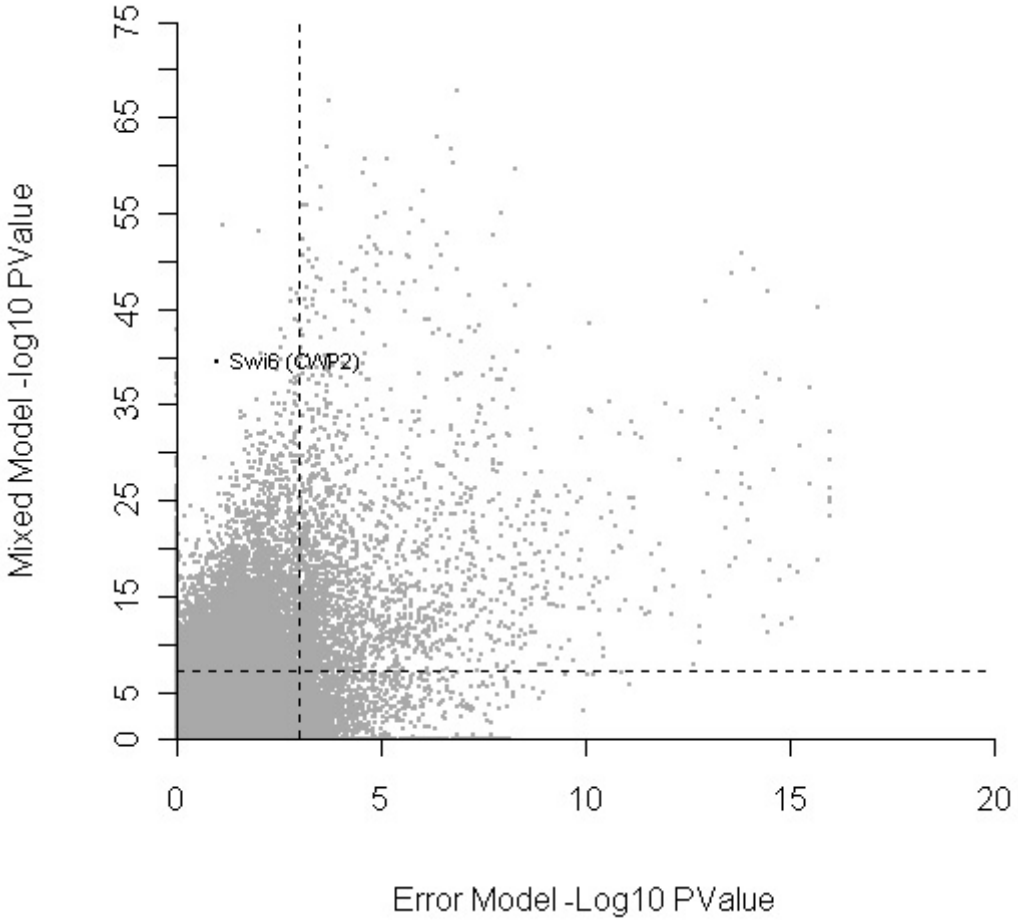

Figure 5. Comparison of negative $\log_{10} p$-values from the mixed model and the error model. The horizontal dashed reference line indicates the Bonferroni cutoff (p-value = 7.5×10^{-8}) used in the mixed model and the vertical dashed reference line indicates the error model cutoff (p-value = 0.001) used in Lee *et al.* (2002). The significant binding of transcription factor Swi6 on *CWP2* promoter identified by the mixed model is indicated in a black, bold point.

Identification of transcriptional regulatory motifs

It has always been of great interest to find out how regulators crosstalk and coordinate in regulating a cellular process. Similar to the method described in

DOI: 10.2202/1544-6115.1045

Shen-Orr *et al*. (2002) and Lee *et al*. (2003), we can draw inference about simple regulatory motifs based on the interactions identified between a transcription factor and its target genes. A regulatory motif represents a compact, modular form that occurs as a pattern in the transcriptional network and is the basic building block of a complex network. Simple regulatory units include autoregulation, feedforward loops, regulator chains, single-input modules (SIM) and multiple-input modules (MIM). We identified these units from the 12,147 interactions found from the mixed models, which included 104 transcription factors and 3,608 genes.

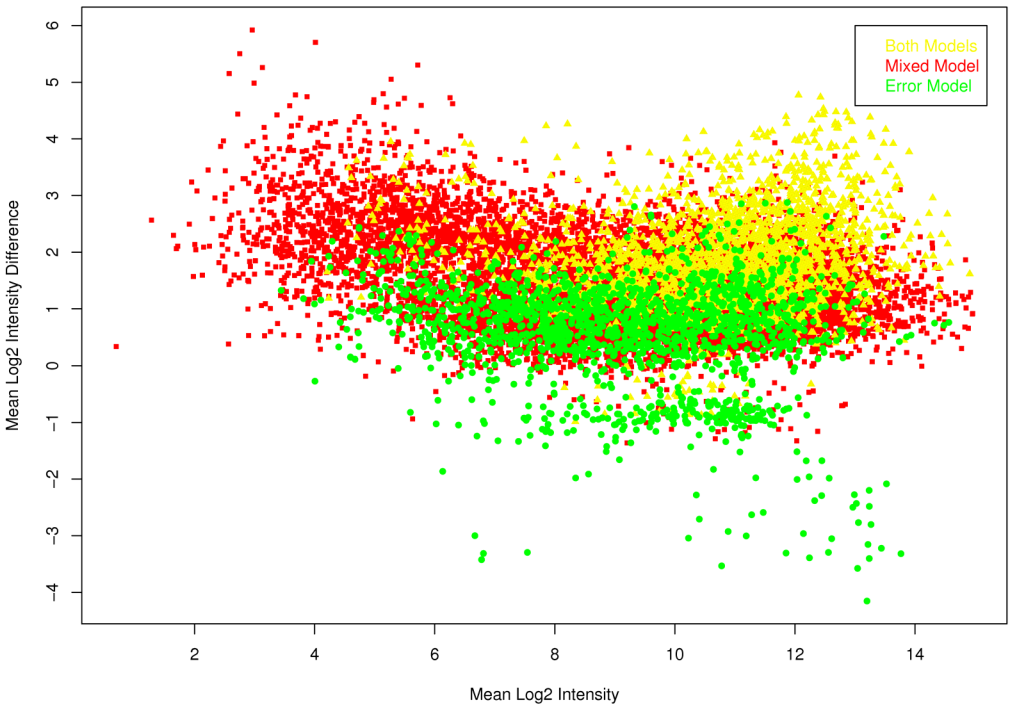

Figure 6. Comparison of "false negatives" from the mixed model and the error model. Analogous to an MA plot, the mean difference of the $\log_2$ intensity between the IP channel and the control channel from the three replicates is plotted against the mean $\log_2$ intensity for three groups: spots found significant by both the mixed model using Bonferroni correction and the error model at 0.001; spots found significant by the mixed model but not the error model and those found significant by the error model but not the mixed model. Note there are a few points below 0 that have positive differences in least squares means between the two probes but the differences of arithmetic means are negative.

In an autoregulation motif, a regulator binds to the promoter of its own gene (Figure 7). Autoregulation serves to amplify cellular responses, reduce the response time to environmental stimuli, and decrease the biosynthetic cost of regulation (Shen-Orr *et al*., 2002). From the 106 transcription factors, we identified fourteen autoregulation motifs, Aro80, Dot6, Gat1, Nrg1, Ino2, Ino4,

Rap1, Rcs1, Smp1, Sok2, Ste12, Swi4, Yap6 and Zap1. Zap1p is a transcription factor that directly controls zinc-responsive gene expression in yeast and is found to increase the expression of the *ZAP1* gene itself, presumably through a positive autoregulatory mechanism. Further support for this autoregulation comes from the identification of a Zap1p binding site within the *ZAP1* gene promoter (Zhao *et al*., 1998). Using the error model and a *p*-value cutoff at 0.001, Lee *et al*. (2003) identified 10 autoregulation motifs, Aro80, Nrg1, Rap1, Rcs1, Smp1, Sok2, Ste12, Swi4, Yap6 and Zap1, all of which are included in the fourteen found by the mixed model. They have also confirmed these 10 binding interactions using conventional chromatin immunoprecipitation.

It has been shown that Reb1p is also an autoregulated transcription factor that binds to its own promoter region (Wang and Warner, 1998). Testing for Reb1p-*REB1* binding yielded a p-value of 7.9×10^{-5}, which failed a multiplicity adjustment using the Bonferroni criterion. Since the Bonferroni correction is a conservative approach to multiple-testing problems, we may risk increasing the false negative rate while minimizing false positives. However, it is noteworthy that the sensitivity of our mixed model is much higher as compared to the error model using the same criterion. As discussed in Wolfinger *et al*. (2001) and Chu *et al*. (2004), this conservative Bonferroni adjustment still serves as a good lower bound to screen for the most significant transcription factors and their target genes for further analysis.

A feedforward loop contains a regulator *X* that regulates a second regulator *Y*, such that *X* and *Y* control a common target gene *G* jointly. A feedforward loop is thought to cause a rapid response to an external signal through amplification of regulators of the target gene. It may also serve as an "AND-gate" control over the target gene *G* when the activation through *X* is transient, and accumulation of both *X* and *Y* signals are required to activate *G* (Shen-Orr *et al*., 2002). We found 316 feedforward loops involving 85 transcription factors, and these potentially control 1,288 genes. For example, Mcm1p binds to the *SWI5* promoter, and both Mcm1p and Swi5p regulate the expression of *CLN3* (Figure 7). During the *S. cerevisiae* cell cycle, the transcriptional regulator Mcm1p is responsible for transcriptional activation of multiple G2-to-M phase specific genes and it has been shown that Mcm1p binds to the promoter region of *SWI5* at sites known to be involved in cell cycle (Althoefer *et al*., 1995). Swi5p activates transcription of genes expressed in G1 phase and at the M/G1 boundary. The target gene *CLN3* encodes a G1 cyclin that allows cells to commit to a new round of division. This forward loop may indicate some kind of temporal control of gene expression that ensures an adequate amount of Cln3p at the start of a new cell cycle. As a comparison, the error model (Lee *et al*., 2002) found Mcm1 and Swi5 form a feedforward loop with each of three target genes *PIR1*, *PIR3* and *YJL160C* but not *CLN3*.

DOI: 10.2202/1544-6115.1045

A regulator chain consists of a chain of regulators in which regulator X_1 binds to the promoter of a second regulator X_2, X_2 then binds to a third one X_3, and so on. We used a recursive algorithm to find regulator chains. For the i^{th} regulator in a chain, we find all regulator promoters that it binds to, place them at the $(i+1)^{th}$ position in the chain, and continue on to the next position until the cascade ends. There are three possible ways to end a chain: 1) the last regulator does not bind any regulator promoter; 2) the last regulator binds to its own promoter; 3) the last regulator binds to a regulator promoter earlier in the chain. A special case in 3) is that the last regulator binds to the promoter of the starting regulator of this chain,

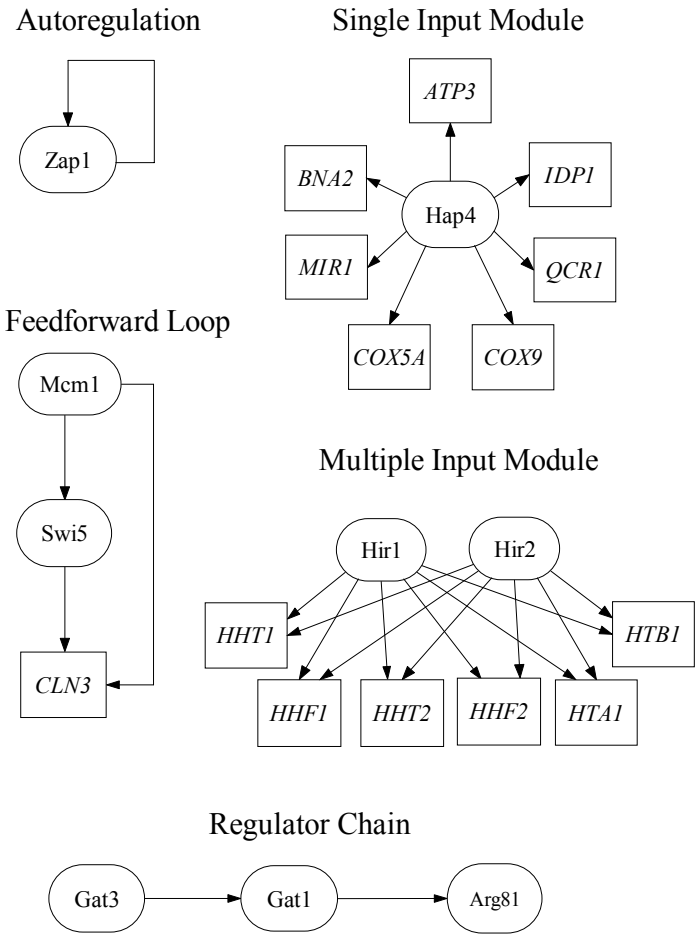

Figure 7. Examples of regulatory motifs. A transcription factor is represented in an eclipse box and a gene is represented in a rectangular box and an arrow indicates binding of a transcription factor to the promoter of its target gene.

resulting in a so-called a multi-component loop as discussed in Lee *et al*. (2002). We identified 1,841 regulator chains from the mixed model results, with lengths varying from 3 to 20. An example of a three-component cascade is Gat3-Gat1-Arg81 in which Gat3p regulates *GAT1* and Gat1p then binds to the promoter of *ARG81* (Figure 7). In *S. cerevisiae*, GAT3 and GAT1 are the GATA-type transcription factors and Gat1p is known to mediate the nitrogen catabolite repression (NCR)-sensitive gene expression (Kuruvilla *et al*., 2001). *ARG81* is a transcription factor required in arginine metabolism. The Gat3-Gat1-Arg81 chain may be indicative of some sequence of transcriptional events that controls nitrogen metabolism. Using the error model (Lee *et al*., 2002), neither the binding of Gat3p to the promoter of *GAT1* nor the binding of Gat1p to that of *ARG81* was identified.

Single-input module (SIM) motifs contain a single regulator X that binds a set of genes G_1, …, G_n, presumably under a specific condition, and these genes have no additional transcriptional regulation. To identify a single-input module, we first found genes whose promoter is bound by only one regulator and then grouped these genes by the regulators. We found 55 SIMs from our 12,147 significant binding interactions, and these potentially control 1,025 target genes. Although these results may not represent all bindings under all conditions in a living cell, a single-input motif still serves as a starting point to look at a coordinated discrete unit of biological function. For example, Hap4p, a global regulator of respiratory gene expression (Forsburg and Guarente, 1989), forms a SIM with seven genes involved in respiration, *ATP3*, *BNA2*, *COX5A*, *COX9*, *MIR1*, *IDP1* and *QCR1* (Figure 7). *ATP3* encodes the gamma subunit of the F1 sector of mitochondrial F1F0 ATP synthase. *BNA2* is required for NAD biosynthesis. *MIR1* product is the mitochondrial phosphate carrier that imports inorganic phosphate into mitochondria. *IDP1* is the gene of mitochondrial NADP-specific isocitrate dehydrogenase that catalyzes the oxidation of isocitrate. *QCR1* encodes the core subunit of the mitochondrial cytochrome bc1 complex. Each of Cox5Ap and Cox9p is a subunit of cytochrome C oxidase, which is the terminal member of the mitochondrial inner membrane electron transport chain. This SIM may also indicate that several genes function stoichiometrically, perhaps because the products assemble together to form a functional protein complex. The results from the error model (Lee *et al*., 2002) indicated that Hap4p forms a SIM with 34 genes including *COX5A*, *IDP1* and *QCR7* but not *ATP3*, *BNA2*, *COX9* or *MIR1*.

Multiple-input module (MIM) motifs consist of a set of regulators that bind together to a common set of genes. A MIM implicates potential coordination of gene expression under various conditions, or a possible coordinated cellular process involving multiple regulators, signal molecules and functioning genes. We identified 826 combinations of two or more regulators that control a set of common target genes. For example, Hir1p and Hir2p, known as negative

DOI: 10.2202/1544-6115.1045

regulators in the transcription of histone genes during the *Saccharomyces cerevisiae* cell cycle, form a MIM with six histone genes, *HTA1*, *HTB1*, *HHT1*, *HHF1*, *HHT2* and *HHF2* (Figure 7). It has been shown that these six histone genes are regulated at the transcriptional level and this transcriptional level regulation plays an important role in the synthesis of the core histones in *Saccharomyces cerevisiae*. Hir1p and Hir2p have been shown to repress the transcription of *HHT1-HHF1* and *HHT2-HHF2*, and the regulation of *HTA1-HTB1* involves another protein Hir3p (Spector *et al*., 1997). These results are consistent with our findings.

To evaluate the statistical significance of observing these network motifs, we compared these results to a completely randomized network. We randomly generated 12,147 interactions among the 106 transcription factors and the 6,279 intergenic regions and recorded the number of motifs for each category. For each motif category, the probability of observing an equal or a greater number of that category is less than 1 instance in 1,000 randomizations except that the observed

Table 2. Multiple-testing adjustment and the effect of *p*-value threshold on the number of network motifs detected. Numbers of network motifs in each category are listed with different *p*-value thresholds using Bonferroni or false discovery rate (FDR) adjustment. The results are ordered by single test *p*-values from the highest to the lowest for the above methods. The total number of tests is 665,574.

Method	Cutoff	AR	FFL	RC	SIM	MIM
Error model	.001	10*	49*	191*, a	90*	295*
FDR	.01	27	985	20,902	48	2,037
FDR	.001	22	677	10,542	53	1,524
FDR	.0001	17	478	8,622	54	1,164
Bonferroni	.05	14	316	1,841	55	826

Note: AR=autoregulation; FFL=feedforward loop; RC=regulator chain; SIM=single-input module; MIM=multiple-input module.

*: result from Lee *et al*. (2002)

a: the number of regulator chain from Lee *et al*. is 188 and they also identified three multi-component loops, which is a special case of regulator chain, so the total number is 191.

number of single input module (SIM) is always 106 in a randomized network, owing to the fact that more genes are likely to be assigned to a unique transcription factor. This indicates that the interactions between transcription factors and genes are highly structured.

Because we used the conservative Bonferroni correction, we expect that the number of true interactions is underestimated, and therefore the number of motifs we identified in each category is also probably an underestimation. For example, if we use an FDR at 0.0001, the number of significant interactions we can identify is 18,678 and the number of motifs in the latter four categories, feedforward loop,

regulator chain, single-input modules, multiple-input modules, will be 478, 8,622, 54, and 1,164, respectively. The number of motifs we found in each category using different *p*-value cutoffs for multiple-testing adjustment is summarized in Table 2.

Discussion

A conventional use of cDNA microarrays is to identify differentially expressed genes. Whether the transcription level of a specific gene is governed by a specific transcription factor is usually inferred indirectly from comparison of certain conditions, for instance, when the transcription factor is deleted or over-expressed. Since transcription typically starts with the association of the transcription factor and co-factors at the promoter site of the target gene, direct study of this protein-DNA interaction has always been of great interest. The introduction of the genome-wide location analysis method (Ren *et al*., 2000) provides a systematic way to infer the binding interactions between a transcription factor and a large number of genes using microarray technology, helping us to understand the mechanisms of transcriptional regulation of these genes.

Analysis of genome-wide location data is similar to that of regular microarrays. Systematic approaches described in Chu *et al*. (2002) provide a general framework for microarray data analysis and can easily be adapted to analyze genome-wide location data. Among the algorithmic steps, data cleaning and outlier detection seems important prior to analysis. In practice, it is not unusual to get low quality arrays since there are many factors that affect the quality of the final image data: quality of clone preparation, uneven DNA printing on the slide, scratches, dust or other artifacts on the array, and non-uniform hybridization, etc. Excluding spots with poor quality in the early stage of microarray analysis is beneficial since the normalization stage usually involves an estimation phase. Several recent studies have addressed performance of quality control during the image-processing phase (Brown *et al*., 2001; Wang *et al*., 2001). When the original images are not available, we can still do some quality control by taking advantage of replicates from the experiment. We showed here that a correlation check between replicates can lead to discovery of poor quality arrays as well as poor quality regions of spots when this check is performed in a block-by-block fashion.

Power to find significant TF-gene interactions was substantially increased, as compared to a *t*-test based error model approach, by fitting a linear mixed model for each gene. Pooling all of the data together to compute variance components leads to identification of the sources of variation affecting the response variable, namely the specific mutant strains that are inferred to have altered binding intensity of a transcription factor with the gene promoter. The error model, by

DOI: 10.2202/1544-6115.1045

taking into consideration the intensity-dependent signal variability that commonly appears in a microarray experiment, estimates the error variance using both a multiplicative and an additive error term, which leads to increased sensitivity and specificity in detecting differential binding interactions. The error model *p*-value for each gene and transcription factor, however, is calculated from only the replicates of that transcription factor, which in this study were only six observations for each gene for each transcription factor. Therefore, the results from the error model are likely to be too conservative and lack power. On the other hand, in our mixed model, all 600 observations for each gene are pooled together, leading to a much more reliable error estimate and increased sensitivity.
An example of the improved resolution of our method is supplied by the report of immunoprecipitations experiments in Iyer *et al.* (2001) that *CWP2*, which encodes a structural protein of the cell wall, is a target gene of Swi6p. Our mixed model identified this binding interaction to be highly significant (*p*-value = 3.9×10^{-40}) whereas the *p*-value from the error model is 0.1 (Figure 5). As seen from Figure 5, the upper left rectangular section contains significant bindings found by the mixed model but not by the error model. If we set the *p*-value cutoff from 0.001 to 0.0001 in the error model (by moving the vertical reference line from 3 to 4), only 594 significant interactions will be retained in the lower right section that are found by the error model but not the mixed model.

It should be noted that unless we are willing to assume there is no gene-specific dye effect, in which case the global dye effect is removed by the normalization model, we need a dye effect in our mixed model that accounts for it. Gene-specific dye effects could arise from differences in the efficiency of fluorophore incorporation and/or differences in hybridization efficiency of the DNA fragments coupled with different dye molecules. In our gene model, we did not include the dye effect because it is completely confounded as a result of the experimental design. This, however, could somehow bias the estimate of our parameter of interest, that is, the transcription factor by probe interaction. Therefore, even with a conservative Bonferroni correction for multiple testing, we expect that the family-wise false positive rate may exceed 0.05 due to this dye bias. It is noteworthy, however, that for this data set, the confounding issue persists no matter which model is used. In order for the dye effect to be estimable, an appropriate experimental design should be adopted that balances the dyes with respect to the probes within each transcription factor. That is, for each transcription factor, an even number of replicates is required in which dye-reversal is performed for the two probes (IP and control). A practical usage of four replicates is thus advocated for this type of study.

It should also be noted that the algorithms described above for finding transcriptional regulatory motifs are solely based on TF-gene interactions. Although they provide a way to possibly identify coordinated regulators and

genes, many other factors, for example intermediate signal transduction molecules along the cellular process pathway that are not directly controlled by the transcription factors, are not likely to be included in the motifs and eventually in assembling transcriptional regulatory networks. Also, the motifs discovered may contain considerable redundancies, especially for regulator chains and multiple-input modules. For example, two regulator chains may have the same start and end transcription factors but with various lengths, or even differ only at one or two positions where the binding path can be substituted with a different transcription factor. Therefore, results from these motifs should be interpreted with care. Additional information, such as extensive gene expression data that can lead to discovery of co-expression of multiple genes, may assist in creation of a clearer picture of transcriptional regulation. A recent study by Bar-Joseph *et al*. (2003) has demonstrated how to combine the genome-wide location data with gene expression profiles to build up the regulatory networks of genetic modules.

Acknowledgements

We would like to thank Paul Magwene, David Orlando, Philip Benfey, and the Systems Biology working group at Duke University for discussions that led to a correction in the normalization model used in a previous version of this paper. The mistake was omitting the channel random effect DA, which led to several genes exhibiting significant differences due plainly to dye bias. This was a clear instance of where biological insight provided a valuable check on statistical conclusions. As follow-on work, we are comparing results in terms of the cluster tuples described in Magwene and Kim (2004).

We thank R.A. Young's lab for publishing their experimental data online, N.J. Rinaldi and B.S. Weir for discussions and comments on the manuscript.

Electronic database

SGD, the *Sacchromyces cerevisiae* genome database: http://yeastgenome.org
Genome-wide location analysis data can be downloaded at:
http://web.wi.mit.edu/young/regulatory_network/

References

Althoefer H, Schleiffer A, Wassmann K, Nordheim A, Ammerer G. 1995. Mcm1 is required to coordinate G2-specific transcription in Sacchromyces cerevisiae. *Mol Cell Biol*. 15: 5917-5928.

DOI: 10.2202/1544-6115.1045

Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert R, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK. 2003. Computational discovery of gene modules and regulatory networks. *Nature Biotech* 21: 1337-1342.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a new and powerful approach to multiple testing. *J Royal Stat Society* 57: 1289-1300.

Brown CS, Goodwin PC, Sorger PK. 2001. Image metrics in the statistical analysis of DNA microarray data. *Proc Natl Acad Sci USA* 98: 8944-8949.

Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG *et al*. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell* 2: 65-73.

Chu TM, Weir BS, Wolfinger RD. 2002. A systematic statistical linear modeling approach to oligonucleotide array experiments. *Math Biosci* 176: 35-51.

Chu TM, Weir BS, Wolfinger RD. 2004. Comparison of Li-Wong and loglinear mixed models for the statistical analysis of oligonucleotide arrays. *Bioinformatics* 20: 500-506.

Chu TM, Wolfinger RD. 2003. Statistical outlier detection for microarray data. *Joint Stat Meeting* 2003.

de Nadal E, Casadome L, Posas F. 2003. Targeting the MEF2-like transcription factor Smp1 by the stress-activated Hog1 mitogen-activated protein kinase. *Mol Cell Biol* 23: 229-237.

Cui X, Hwang JTG, Qiu J, Blades NJ, Churchill GA. 2005. Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics* 6: 59-75.

DeRisi JL, Lyer VR, Brown PO. 1997. Exploring the metabolic and genetic control of gene expression o a genomic scale. *Science* 278: 680-686.

Forsburg SL, Guarente L. 1989. Identification and characterization of HAP4: a third component of the CCAAT-bound HAP2/HAP3 heteromer. *Genes Dev* 3: 1166-1178.

Feng S, Wolfinger RD, Chu TM, Gibson G, McGraw LA. 2004. Empirical Bayesian analysis of variance component models for microarray data. (Submitted)

Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Bostein D, Brown PO. 2000. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell* 11: 4241-4257.

Iyer VR, Horak CE, Scafe CS, Bostein D, Snyder M, Brown PO. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. *Nature* 409: 533 – 538.

Kerr MK, Martin GA, Churchill GA. 2000. Analysis of variance for gene expression microarray data. *J Comp Biol* 7: 819-837.

Kuruvilla KG, Shamji AF, Schreiber SL. 2001. Carbon- and nitrogen-quality signaling to translation are mediated by distinct GATA-type transcription factors. *Proc Natl Acad Sci USA* 98: 7283-7288.

Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM *et al*. 2002. Transcriptional regulatory network in Sacchromyces cerevisiae. *Science* 298: 799-804.
Lönnstedt I, Speed TP (2002) Replicated Microarray Data. *Stat Sinica* 12: 31-46.
Orlando V. 2000. Mapping chromosomal proteins *in vivo* by formaldehyde-crosslinked-chromatin immunoprecipitation. *Trends Biochem Sci* 25: 99-104.
Magwene PM, Kim J. (2004) Estimating genomic coexpression networks using first-order conditional independence. *Genome Biol* 5: R100.
Pe'er D, Regev A, Elidan G, Friedman N. 2001. Inferring subnetworks from perturbed expression profiles. *Bioinformatics* 17: S215-S224.
Ren B, Robert F, Wyrick JJ, Aparico O, Jennings EG, Simon I, Zeitlinger J, *et al*. 2000. Genome-wide location and function of DNA binding proteins. *Science* 290: 2306-2309.
Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, Bennett HA, He YD *et al*. 2000. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* 287: 873-880.
Segal E, Taskar B, Gasch A, Friedman N, Koller D. 2001. Rich probabilistic models for gene expression. *Bioinformatics* 17: S243-S252.
Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N. 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genet* 34: 166-176.
Shen-Orr SS, Milo R, Mangan S, Alon U. 2002. Network motifs in the transcriptional reglation of *Escherichia coli*. *Nature Genet* 31: 65-68.
Spector MS, Raff A, DeSilva H, Lee K, Osley MA. 1997. Hir1p and Hir2p function as transcriptional corepressors to regulate histone gene transcription in the *Saccharomyces cerevisiae* cell cycle. *Mol Cell Biol* 17: 545-552.
Stoughton R, Dai H. 2002. Statistical combining of cell expression profiles. US Patent 6351712.
Wang KL, Warner JR. 1998. Positive and negative autoregulation of REB1 transcription in *Sacchromyces cerevisiae*. *Mol Cell Biol* 18: 4368-4376.
Wang X, Ghosh S, Guo SW. 2001. Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res* 29: E75.
Ward MP, Gimeno CJ, Fink GR, Garrett S. 1995. SOK2 may regulate cyclic AMP-dependent protein kinase-stimulated growth and pseudohyphal development by repressing transcription. *Mol Cell Biol* 15: 6857-6863.
Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. 2001. Assessing gene significance from cDNA microarray expression data via mixed models. *J Comp Biol* 8: 625-637.
Zhao H, Butler E, Rodgers J, Spizzo T, duesterhoeft S, Eide D. 1998. Regulation of Zinc Homeostasis in Yeast by Binding of the ZAP1 Transcriptional Activator to Zinc-responsive Promoter Elements. *J Biol Chem* 273: 28713-28720.

DOI: 10.2202/1544-6115.1045